UNTANGLING ATTRIBUTION: WHY IT HAS FAILED AND HOW IT SHOULD BE DONE

MANOS ANTONAKAKIS
SEPTEMBER 27, 2017
WHAT IS AND WHY SHOULD WE CARE ABOUT ATTRIBUTION?

• What is attack attribution?
  • It is the process through which you can link illicit cyber activities with a virtual (or physical) entity.

• Why should we care?
  • **As Governments**, to protect your critical infrastructure and the nation’s democracy
  • **As large organizations**, to be aware of the groups (and how unique they are) that coming after your assets
  • **As individuals**, so your law enforcement can lock behind bars the common criminal coming after your data (financial or otherwise)
HOW WE DO ATTRIBUTION TODAY?

In the best case scenario, we are confused.

In the worst case scenario we conduct “attribution operations” so we can name groups after (cute) animals and other marketing shenanigans.
BRINGING RIGOR INTO ATTRIBUTION OPERATIONS

• Attribution is a continuous process
  • It has at least 4 phases, and you need data from at least the first 3
  • Certain phases can give you certain insights
  • The “Delta” between these phases is very important to the success of the attribution operation
  • *We CANNOT be waiting for the malware!!!*

• For example
  • If t0 << t1, you might not be able to comment on the reconnaissance
  • If t1 << t2, you might be looking at multiple attack groups
  • If t2 << t3, you might be looking at different classes of attackers or mix-attacks (i.e., non-targeted)
  • If t3 ~= t4, you might be looking into a lot of noise
CONDUCTING ATTRIBUTION AT SCALE

**A B C D: Always Be Collecting Data**
- Start with your own data first.
- Next, the Internet is full of interesting datasets.
- Many organizations are willing to share privacy preserving data for securing the Internet.

**Know & Mine Your Data**
- You need to be able to store and marshal your data efficiently.
- You need to be able to mine the data at scale, utilizing machine learning.
MYTH VS. REALITY
BURSTING THE MACHINE LEARNING BUBBLE

• You know **you are not** doing (or need) machine learning when,
  • your marketing team is putting random (yet fancy) words from machine learning textbooks together for your products.
  • you have more executives than ML experts working with threat researchers.
  • you are using a less than 100TB and a dataset that is more than one-year old to run your model.
  • you have yet to understand the security intuition behind the features you decided to compute.
  • you cannot detect anything more than a signature-based IDS.

• You know **you are** doing (and actually need) machine learning when,
  • the data you want to work with need a greater than 50 node Hadoop cluster.
  • the first thing you have to do is unsupervised operations to help you understand the signal.
  • relational combining of datasets is the first thought in the morning and the last thought before you go to bed.
  • adversarial learning is part of your detection modeling decisions.
  • **you are discovering things that others do not on a regular basis. (The most important ML property at the end of the day ... )**
This is the first time in history that we can stack the deck to our favor!

Attribution is the game of sorting through mistakes. For once, we are not behind the eight ball.
THANK YOU!

ASTROLAVOS LAB
ASTROLAVOS.GATECH.EDU

MANOS ANTONAKAKIS
MANOS@GATECH.EDU